Beyond Validation

Using programmed diagnostics to learn about, monitor, and successfully complete your DH project

Martin Holmes and Joey Takeda
University of Victoria Endings Project

The Endings Project

- How to complete, publish and walk away from your digital edition project…

University of Victoria

The Endings Project

- How to complete, publish and walk away from your digital edition project…
- ...and have it last for 50 years.

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada
University of Victoria

The Projects

- The Map of Early Modern London ↗
- Le Mariage sous l'ancien régime ↗
- The Nxaʔamxcín Dictionary Database
- The Robert Graves Diary ↗
- The Scandinavian-Canadian Studies Journal ↗

University of Victoria

Validation is great

- ...especially if you have created a solid, tightly-constrained schema.
- But it’s not enough...

University of Victoria

Extra Schematron

is a great help. We use Schematron to:

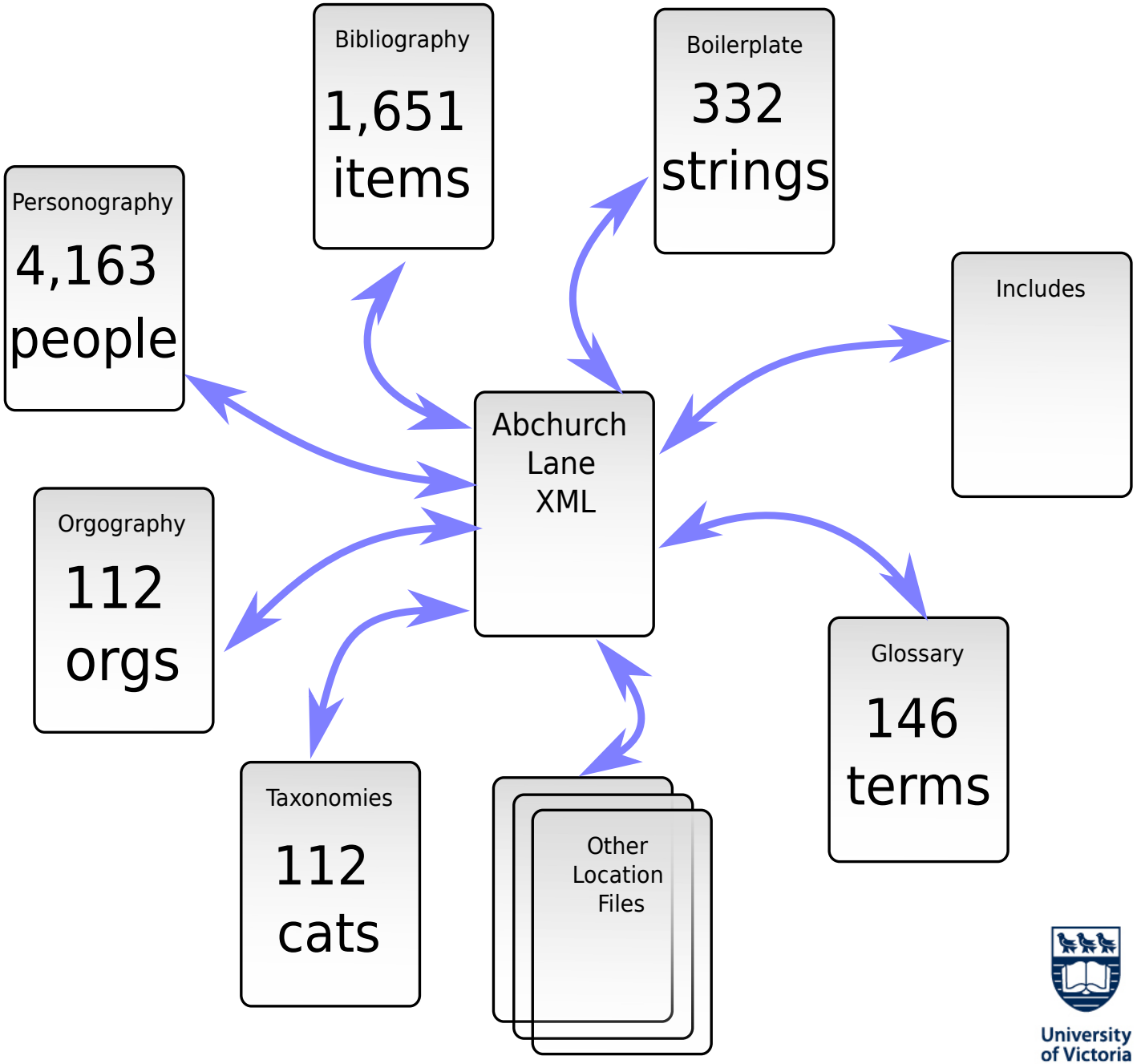
- enforce curly apostrophes in text
- ban quotation marks (use <quote> etc.)
- ban leading and trailing spaces in some tags
- enforce correct capitalization
- limit the length of abstracts
- ...and lots more...

University of Victoria

But that's not enough

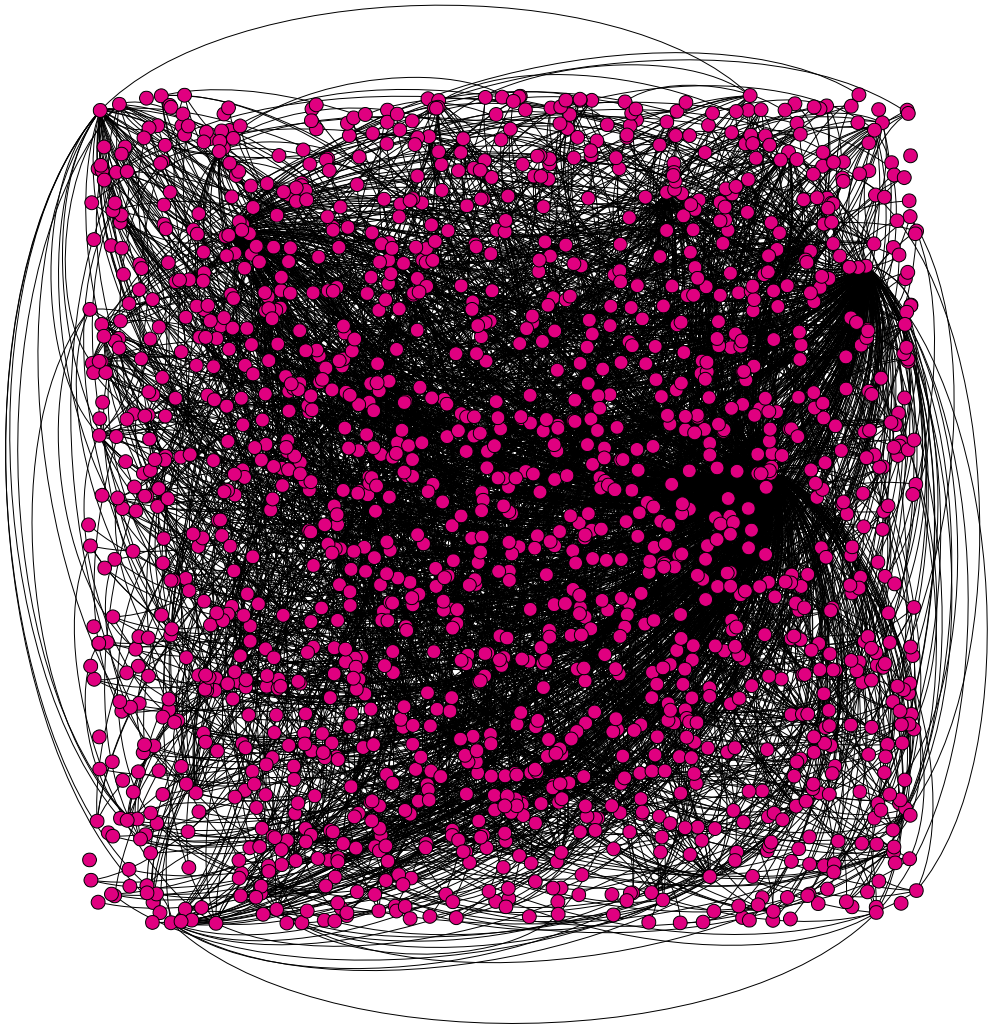
Validation works at the document level, but a digital edition project is typically a large collection of interrelated resources.

University of Victoria



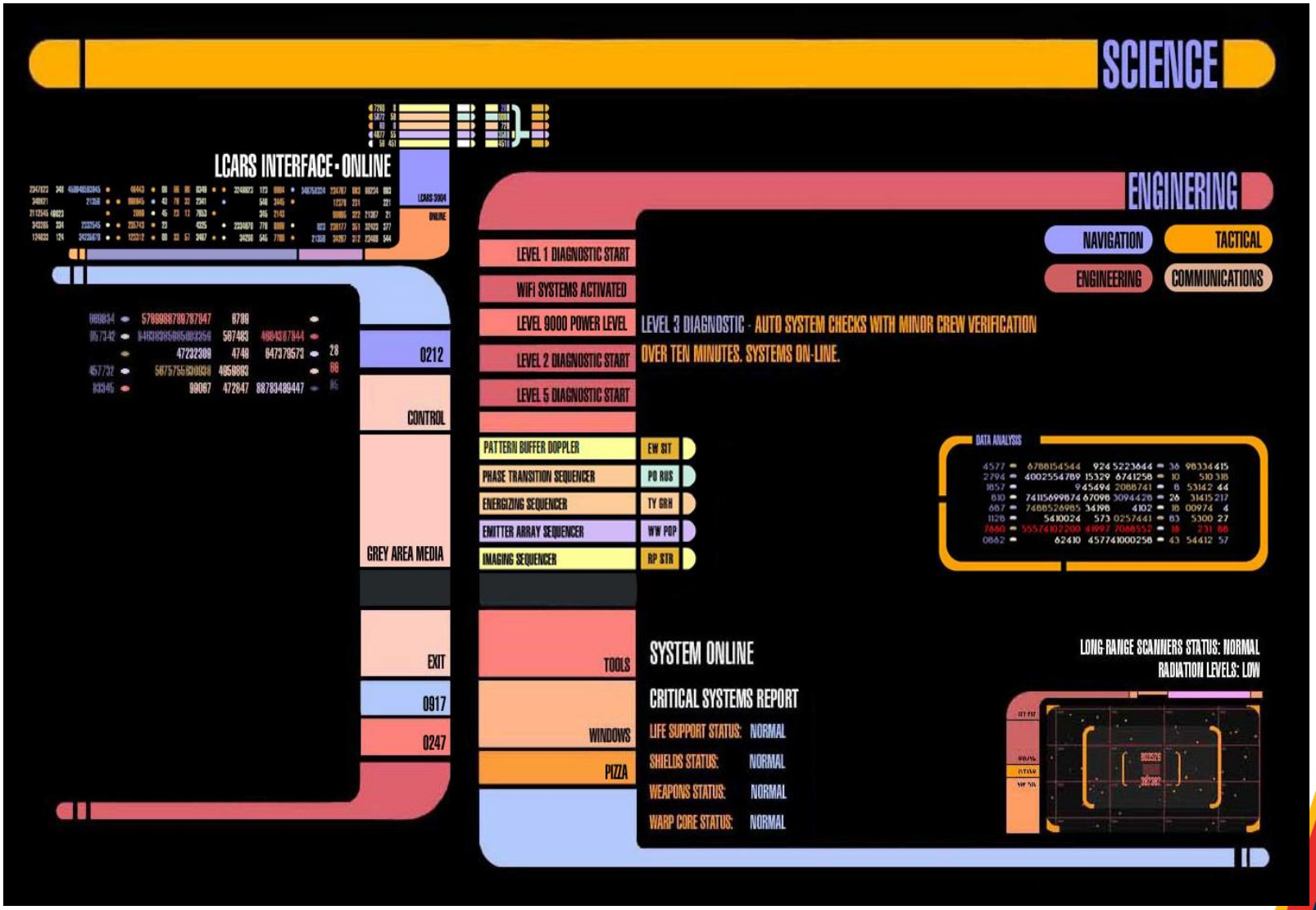
University of Victoria

Network graph



University of Victoria

Diagnostics to the rescue!



DIAGNOSTICS

LEVEL 1

COHERENCE and **CONSISTENCY**

DIAGNOSTICS

LEVEL 1

COHERENCE and **CONSISTENCY**

Project-level checking of referential integrity:

DIAGNOSTICS

LEVEL 1

COHERENCE and **CONSISTENCY**

Project-level checking of referential integrity:

- non-existent pointers/targets

DIAGNOSTICS

LEVEL 1

COHERENCE and **CONSISTENCY**

Project-level checking of referential integrity:

- non-existent pointers/targets
- duplicate @xml:ids

DIAGNOSTICS

LEVEL 1

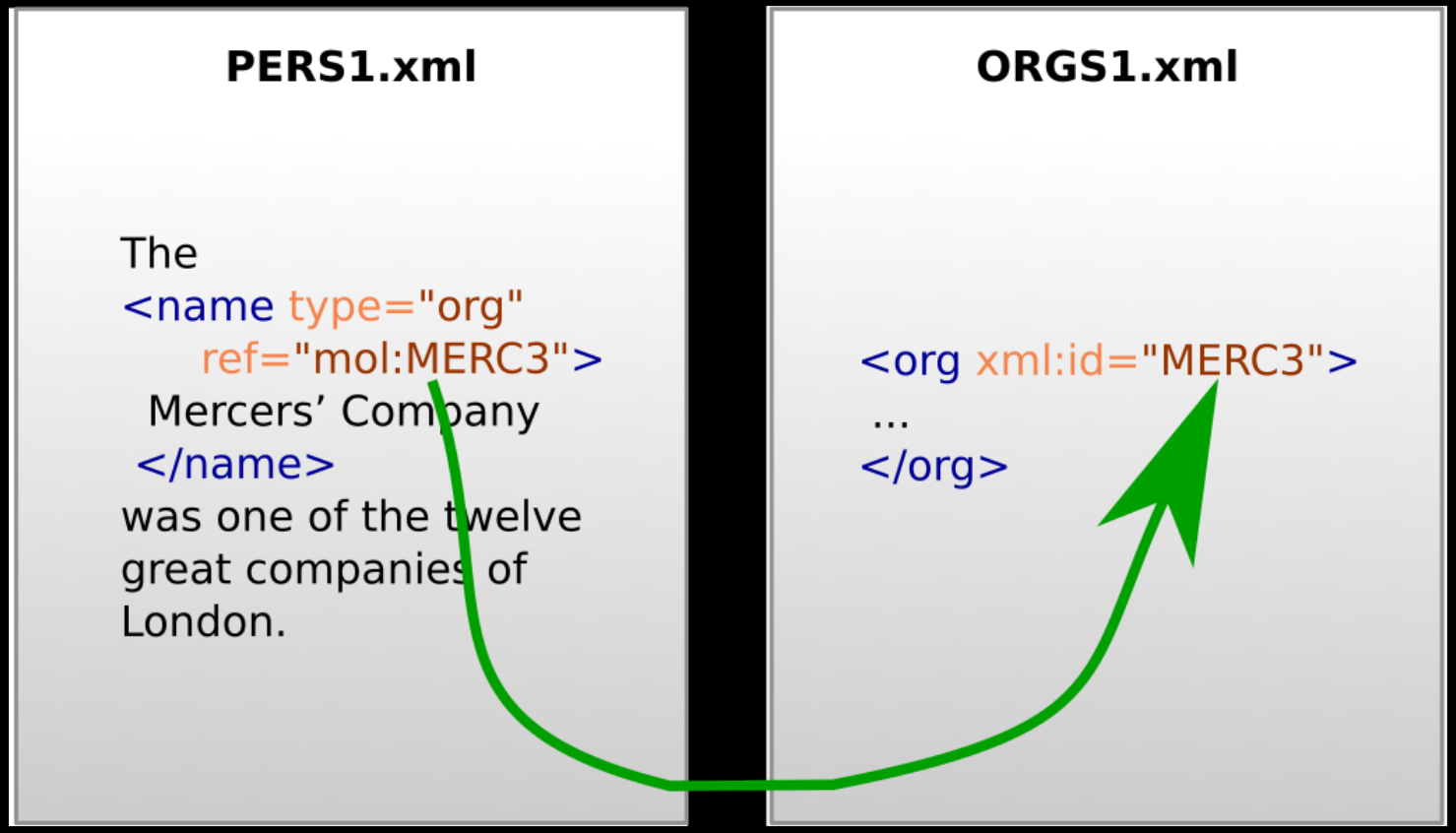
COHERENCE and **CONSISTENCY**

Project-level checking of referential integrity:

- non-existent pointers/targets
- duplicate @xml:ids
- entity type errors (persName / placeName / orgName)

DIAGNOSTICS

LEVEL 1



- Every <name type="org"> points at an @xml:id which exists in the project.
- The element pointed at by <name type="org"> is an <org> element in the ORGS1.xml document.
- Every <name> element which points at an <org> element in ORGS1.xml has @type="org".

University of Victoria

- Every <name type="org"> points at an @xml:id which exists in the project.
- The element pointed at by <name type="org"> is an <org> element in the ORGS1.xml document.
- Every <name> element which points at an <org> element in ORGS1.xml has @type="org".

  - Can Schematron do this?

University of Victoria

- Every <name type="org"> points at an @xml:id which exists in the project.
- The element pointed at by <name type="org"> is an <org> element in the ORGS1.xml document.
- Every <name> element which points at an <org> element in ORGS1.xml has @type="org".

  - Can Schematron do this?
  - Yes, but only for small projects.

University of Victoria

DIAGNOSTICS

LEVEL 2

COMPLETENESS

DIAGNOSTICS

LEVEL 2

COMPLETENESS

Generate to-do lists, identify possible errors requiring human judgement:

DIAGNOSTICS

LEVEL 2

COMPLETENESS

Generate to-do lists, identify possible errors requiring human judgement:

- Suspiciously similar bibliography entries

DIAGNOSTICS

LEVEL 2

COMPLETENESS

Generate to-do lists, identify possible errors requiring human judgement:

- Suspiciously similar bibliography entries
- different names tagged to point to the same <person>

DIAGNOSTICS

LEVEL 2

COMPLETENESS

Generate to-do lists, identify possible errors requiring human judgement:

- Suspiciously similar bibliography entries
- different names tagged to point to the same <person>
- unreachable (unlinked) documents

Possible duplicate <bibl> entries (24) ▼

Explanation

These bibliography entries appear very similar, as measured by a similarity metric, so they are possibly duplicates.

- *Closer to 1 indicates higher similarity.*
- *Closer to 0 indicates lower similarity.*

- <bibl> DEKK2 appears similar to SMAL1 (similarity score 0.925):

  Dekker, Thomas. The Shoemaker's Holiday. Ed. R.L. Smallwood and Stanley Wells. Manchester: Manchester UP, 1979. The Revels Plays.

  Smallwood, R.L., and Stanley Wells, eds. The Shoemaker's Holiday. By Thomas Dekker. Manchester: Manchester UP, 1979. The Revels Plays.

- <bibl> DEKK13 appears similar to SMUT1 (similarity score 0.90697676):

  Dekker, Thomas, Stephen Harrison, Ben Jonson, and Thomas Middleton. The Whole Royal and Magnificent Entertainment of King James through the City of London, 15 March 1604, with the Arches of Triumph. Ed. R. Malcolm Smuts. Thomas Middleton: The Collected Works. Gen. ed. Gary Taylor and John Lavagnino. Oxford: Oxford UP, 2007. 219-79.

  Smuts, R. Malcolm, ed. The Whole Royal and Magnificent Entertainment of King James through the City of London, 15 March 1604, with the Arches of Triumph. Thomas Middleton: The Collected Works. Gen. ed. Gary Taylor and John Lavagnino. Oxford: Oxford UP, 2007.

Beware

- Diagnostics can get out of hand.
- The Moses Dictionary Project has gone nuts with their diagnostics.
- Possible result: project paralysis.

University of Victoria

DIAGNOSTICS

LEVEL 3

PROGRESS AND PREDICTIONS

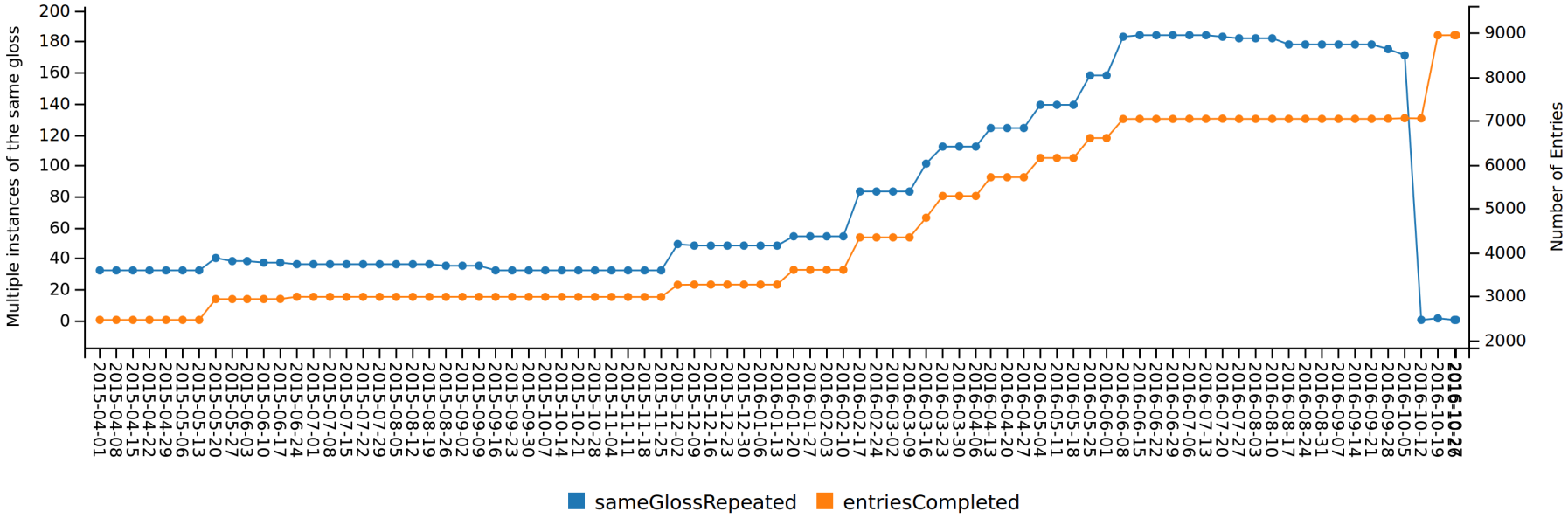
DIAGNOSTICS

LEVEL 3

PROGRESS AND PREDICTIONS

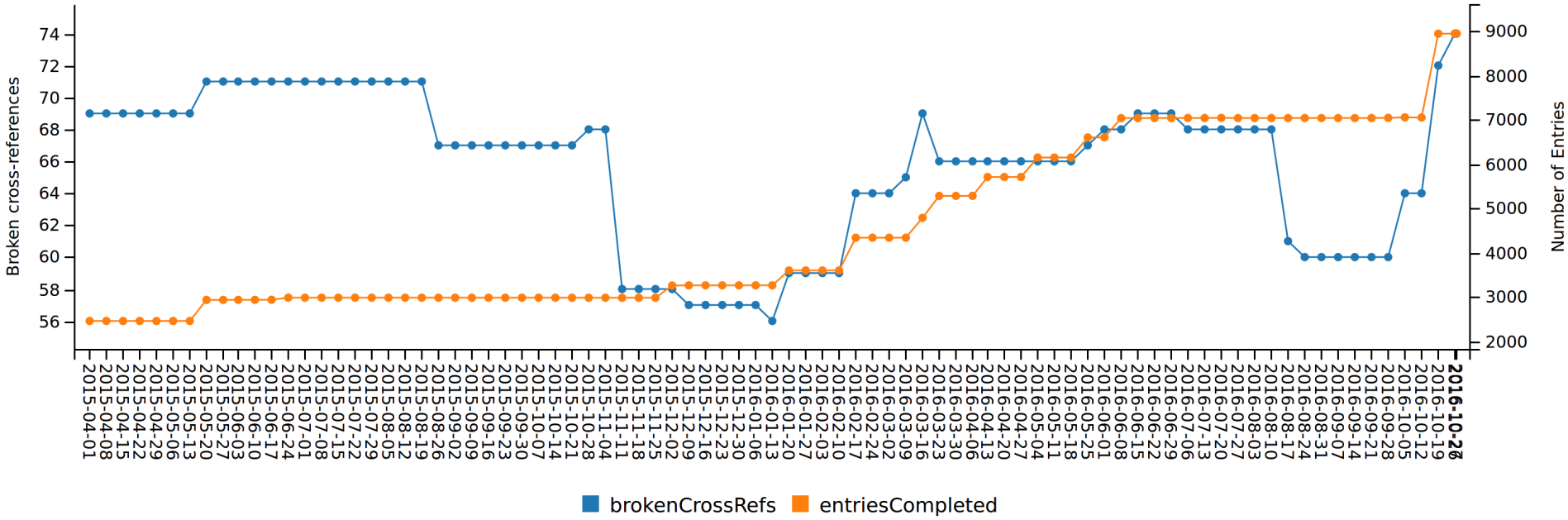
Provide progress analysis: are we hitting milestones?

How are we doing?



Duplicate glosses rise with new entries, until an editing blitz fixes them.

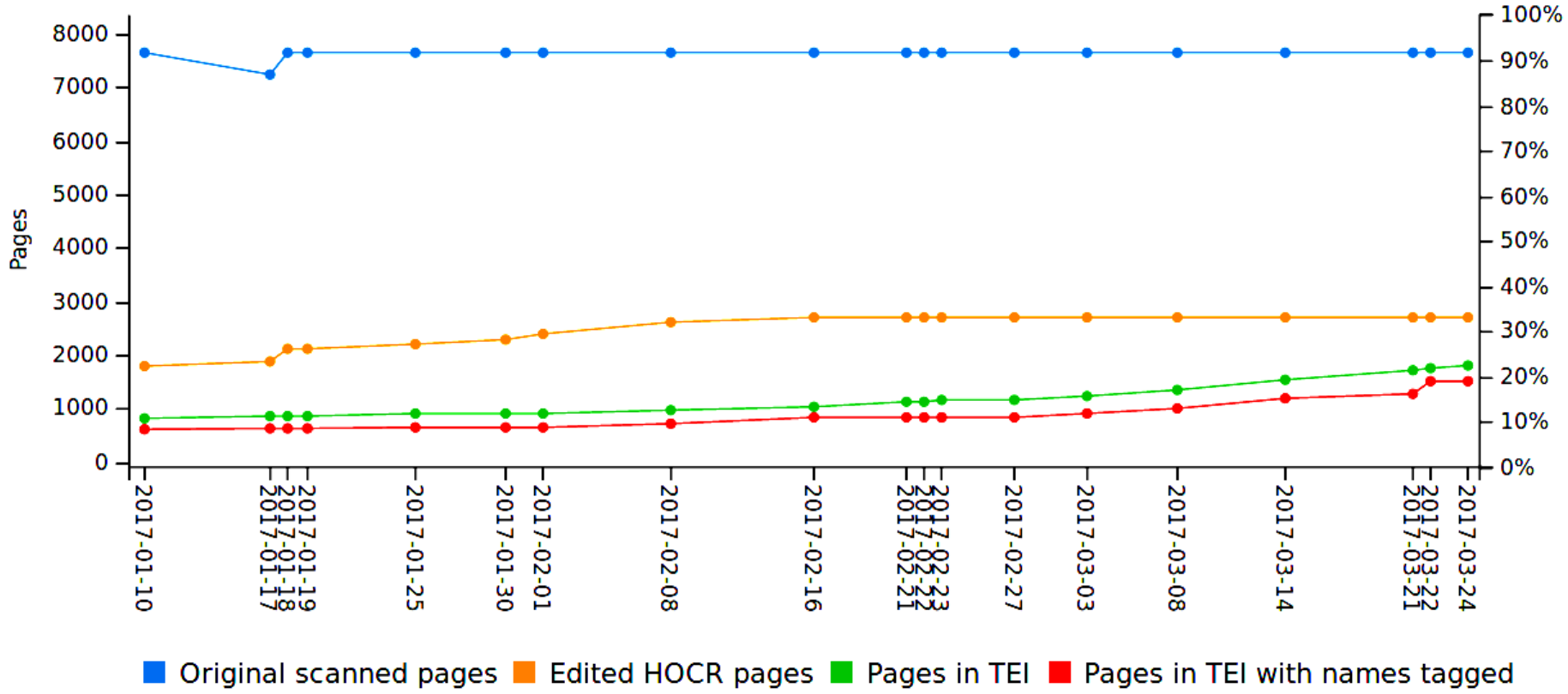
University of Victoria



Broken cross-references keep coming back.

University of Victoria

The Confederation Debates: Progress Chart



Total names tagged so far: 5626

Problematic "unspecified" names: 129

Projected completion dates

- edited HOCR pages: 2017-08-04
- pages in TEI: 2017-11-16
- fully-name-tagged pages in TEI: 2018-01-15

A basic diagnostic toolkit

https://github.com/projectEndings/diagnostics

University of Victoria

- Run it against a folder containing a TEI project.
- It checks that:
  - All pointer attributes within a document point to @xml:ids that exist in the document.
  - All pointers to other documents in the collection, or to @xml:id attributes in those documents, are correct.
  - All values for the @xml:lang attribute are correct.

University of Victoria

DIAGNOSTICS

PASSED

ALL TESTS SUCCESSFULLY COMPLETED